THE DATAFETCH LIBRARY OF FUNCTIONS FOR THE RETRIEVAL AND INTERPRETATION OF THERMOPHYSICAL DATA FROM THE TRC SOURCE DATABASE

Randolph C. Wilhoit*

Abstract

The DataFetch library is a collection of functions that may be compiled and linked to user-written application programs. They extract information from a local version of the TRC SOURCE Database. The database is an extensive archive of experimental values of thermodynamic, thermochemical, and transport properties of pure compounds, chemical reactions, and binary and ternary mixtures. Application programs may call the library functions with parameters that specify the kind of data to be retrieved. The results are returned in a series of memory-resident buffers. The library functions operate at a series of processing levels. The lowest level returns the direct experimental values, along with associated information that identifies the components, properties, phases, literature references, sample descriptions, estimated uncertainties, etc. Higher levels return groups of related properties, normalized values of properties, selections of the most accurate values, and fits to smoothing functions. The library functions are written in C++ and interact with a local version of the SOURCE database. They are compatible with any platform that supports a C++ compiler in either a stand-alone or client-server mode.

KEY WORDS: database, thermophysical properties, data interpretation, application programs, library functions

*Thermodynamics Research Center, Texas A&M University, College Station TX 77843

1 INTRODUCTION

In the Thirteenth Symposium on Thermophysical Properties we addressed barriers to the efficient utilization of thermophysical property data [1]. As the amount of experimental data accumulates in the world's literature it becomes increasingly more difficult and expensive to recover and them. During the past two centuries this task has been eased by gathering data from scattered reports into compact compilations. These have taken the form of review articles published in journals, monographs, books, or sets of books. In recent years they often appear as computer readable databases. Some are issued as one-time publications and some as a continuing series of reports which attempt to keep up with current research. The kind of coverage, the method of presentation, and the extent of interpretation of the data varies widely among existing compilations

Irrespective of the form of the compilation, those which reflect the state of knowledge of a subject at a particular time are static. In order to keep up with the world's accumulation of data they must be periodically updated and re-issued. The need to re-work much of the same data for each revision and the difficulty in anticipating in advance what users may want causes wasted effort.

In our presentation two years ago we described dynamic compilations. These are produced to order by the user at the time of need. They require two components. One is a continually updated computer readable archive of experimental data. The other is software that interacts with the archive, locates data of interest, interprets them, and converts them into a usable form. At that time we described the SOURCE database, created and maintained by the Thermodynamics Research Center, which is suitable as

an archive. The SOURCE database has since been improved and expanded in size.

Now we will describe software to support dynamic compilations. However this line of thought has been carried further. Most users of thermophysical properties do not regard the numbers themselves as the final objective. Rather, they use these numbers to accomplish some further goal. Practical goals include the design and operation of manufacturing or transportation facilities, evaluation of new manufacturing processes, or improvement of health, safety, and environmental quality. Academic goals include studies of molecular structure-property relationships and testing of theories of matter.

In recent times data applications, especially if they require large amounts of numerical data, are made with the help of computer programs. Call these application programs. Since the data are only an intermediate step in this process they do not need to exist in a human readable form at all. It is more efficient and convenient for application programs to access needed data directly. For this, application programs should be able to link directly to the software component of the dynamic compilation process.

2 GENERAL SPECIFICATIONS OF THE DATAFETCH LIBRARY FUNCTIONS

The DataFetch Library is a collection of computer functions which support dynamic compilations and direct linkage to application programs. They are distributed in object form which may be linked to application programs during compilation. When an application program calls a DataFetch function it passes parameters which specify the data to be retrieved. Results are returned in memoryr-resident buffers which may be accessed by the application program. The library can exist in either a static or dynamically linked

form.

The SOURCE database contains primarily directly measured values of properties of systems of specified composition. Sometimes smoothed or derived values are included, especially if the direct experimental values are not reported. It consists of 35 tables. Some store Registry Numbers[1] and compound names and formulas. Other tables contain literature references and author names. Tables for storing properties of pure compounds, binary mixtures, ternary mixtures, and calorimetric and equilibrium constants for chemical reactions are present. All kinds of thermodynamic and thermochemical properties of all phases and transport properties of fluid are accommodated. Pure compounds and components of mixtures and reactions are identified by Registry Numbers. Data tables include values of state variables, properties, and estimated uncertainties. One table contains descriptions of samples used in the measurements. The database also includes a variety of metadata which describe the method of presentation, units in the original document, and other information to help in the evaluation and selection. The tables are indexed and linked so that related data may be retrieved. Properties and phases are identified by formally assigned codes. Detailed documentation of the SOURCE database is available [2]. Reference [3] gives an overview.

DataFetch functions that return data operate at a series of processing levels.

Level 1: Returns data for a requested property and system (pure, mixture or chemical reaction) from a particular database table. These contain information directly extracted from the database with little change.

[1]Numbers assigned by Chemical Abstracts are used whenever possible, otherwise numbers are assigned by TRC

Level 2: Returns values of closely related groups of properties extracted from one or more relevant database tables.

Level 3: Returns data for a property group in a normalized form.

Level 4: Returns selected values. Normalized data from level 3, which can now be inter-compared are screened to identify the best (more accurate) values.

Level 5: Returns parameters of smoothing functions fit to the output of Level 4 for a group of properties by the least squares criteria, along with statistical measures of goodness of fit.

Level 6: Returns parameters of functions for calculating internally consistent properties. These satisfy thermodynamic constraints among properties. It operates on combinations of several Level 4 results.

Level 7: Carries out the same processing as Levels 5 and 6, but accepts a combination of Level 4 output with data from theoretical calculations or empirical correlations.

Functions at each level except the first use results returned from lower levels. Functions at Levels 1-4 return the requested numerical data, as well as associated information such as identification of properties and phases, sample descriptions, literature references, and available metadata. They closely reflect the organization of data in the SOURCE database.

3 ORGANIZATION OF THE SOURCE DATABASE

Database tables that store numerical values of thermophysical properties are defined by the number of components in the system and the number of associated independent state variables. The total number of state variables equals the degree of freedom of the system as calculated by the Gibbs' phase rule, $F = C - P + 2$. C is the number of independent components in the system and P is the number of phases. The database tables are based on the "effective" degree of freedom, which may be less than that calculated by the Gibbs' phase rule. For example if one or more state variables are kept constant for a set of data they are not counted in the "effective" degree of freedom. The effective degree of freedom is also reduced by one for certain special states such as liquid-gas or liquid-liquid critical states and azeotropic states. Although the Gibbs' phase rule does not apply to transport properties (viscosity, thermal conductivity, diffusivity) the concept of effective degrees of freedom is applied to these properties as well.

Data with zero effective degrees of freedom are stored in tables which include all the descriptive metadata as well as the property values and uncertainties.

Data with more that zero effective degrees of freedom are stored in a pair of database tables. A header table contains information which describes the data. These include codes for the property and the state variables, an identification of phases, sample numbers, and other metadata which describe the way the data was presented in the original document and which help in the evaluation. The value of any state variable which is kept constant for a data set is also stored in the header table. The numerical values of state variables not kept constant and the property are stored in a separate data table

along with an uncertainty value for each property.

Both kinds of tables are indexed by registry numbers of components, a key for the literature reference, the property code, and a set number. A header table and its associated data table stores the value of only one kind of property.

Calorimetric heats of chemical reaction are treated as having zero degrees of freedom and stored in one table. Most equilibrium constant data are considered as having one degree of freedom where temperature is usually the state variable. They are stored in a header-data pair of tables. These thermochemical property tables are indexed by a reaction classification code and by registry numbers of four reaction participants. Registry numbers for any additional reaction participants as well as coefficients in the balanced chemical equations are stored in the header tables.

In principle the distinction between a property and the state variables is arbitrary. For example the pressure-temperature pair that defines an equilibrium between the liquid and vapor phases of a pure component may be called a boiling point, where the temperature is the property and the pressure is the state variable, or a vapor pressure, where pressure is the property and temperature the variable. Another example is P-V-T data for a single phase system of a pure component. Either pressure (P), volume (V), or temperature (T) may be chosen as the property and the other two would be the state variables. In a set of isothermal P-V-T data temperature is kept constant. The value of the constant temperature is placed in the header record and the data values have one effective degree of freedom. In a set of isobaric P-V-T data the pressure is constant, and in isochoric P-V-T data the volume is constant. If nothing is kept constant the property has two degrees of freedom. Sets of vapor-liquid equilibrium data in a binary system

may be characterized by values of pressure, temperature, and composition of liquid and vapor phases (p,t,x,y data). The Gibbs' phase rule gives two degrees of freedom for this system. Thus any two of these values may be chosen as the state variables, and another as the property. Since there are four variables the system is characterized by two properties which are stored in different data sets in the database. If one of these variables is kept constant for a set of data it then has one effective degree of freedom. Sets of data in which only pressure, temperature, and liquid composition are measured (p,t,x data) have only one property.

Any choice of property and state variables for a system may be accommodated in the SOURCE database. Generally the choice reflects the way the data was presented in the original document. However this still leaves some variation in the way data is stored in the database. In any case properties and state variables are clearly identified in the header records for each data set.

4 FURTHER DESCRIPTION OF PROCESSING LEVELS

Level 1 DataFetch retrievals extract data from the database as stored. The information is returned in five blocks. These are:

1. Descriptive information from the header records
2. Numerical values of state variables and properties from the data records
3. Description of reference states, if any are used, extracted from header records
4. Description of samples used in the measurements

5. Literature references to original sources of the data

The records in these blocks are connected by identification numbers. A Level 1 function exists for each table with zero degrees of freedom and for each header-data pair of tables with one or more effective degrees of freedom. Parameters passed to Level 1 functions include registry numbers for the components and one or more codes that specify property-phase combinations.

Level 2 functions return data in a manner similar to Level 1. However Level 2 combines closely related properties in one set of return blocks. These properties may be extracted from more than one header-data table pair. Thus for example the the P-T group of liquid-vapor phases of pure compounds would return data labeled both as boiling point and vapor pressure. These data for reside in three different table combinations in the database. Similarly the P-V-T group for single phases of pure compounds combines all data irrespective of which variable was chosen as the property. The vapor-liquid group for binary systems would return all data involving pressure-temperature-composition variables for binary mixtures. Parameters for Level 2 functions include registry numbers and a code for the property group.

Level 3 normalizes the data returned from a Level 2 retrieval. Normalization makes a standard choice of property and state variables and converts them to a consistent form. The uncertainties listed for observed properties in the database are also propagated to the normalized form. The precise meaning of normalization depends somewhat on the property group processed by Level 2. Most properties that have zero degrees of freedom do not require normalization. Examples are critical temperature and pressure, triple point temperatures, and normal boiling points of pure compounds. The vapor pressure-

boiling point group for pure components are presented with pressure as the property and temperature as the state variable.

The normalized results for the one-phase single component for P-V-T data group from Level 2 are presented with volume as the property and temperature and pressure, in that order, as the state variables. There are several kinds of volumetric properties, such as specific density, molar density, specific volume, molar volume, and compressibility factor. Level 3 converts all these to one kind of property, say specific density.

Normalization of data for mixtures is more complicated. For example volumetric properties of mixtures, in addition to those listed above, include excess volume, partial molal volume, and apparent molal volume. Compositions of mixtures may be expressed as mole fraction, mass fraction, volume fraction, molality, molarity, etc. Normalization converts all of these into one kind of composition variable.

Normalization is a critical step that is required for all higher levels of data handling. Normalization may require complicated conversions and reorganization. It may also require auxiliary data, not part of the property group being normalized. The auxiliary data may be obtained from a separate Level 5 retrieval for the auxiliary properties, or may obtained from the parameter database discussed below. Propagation of uncertainties into the final choice of property requires a knowledge of derivatives of the property with respect to state variables. It may be necessary to carry out a Level 5 calculation using some nominal uncertainty estimates to obtain the derivatives. This sets up a loop which should converge after a few iterations.

Level 4 operates on data returned from Level 3. It selects the best (most accurate) values of properties whenever duplications, or near duplications, exist among the re-

ported properties. This screening is based primarily on the estimated uncertainties of properties, either given directly in the database or propagated to the normalization data.

The selection of data with zero effective degrees of freedom is simple. It is only necessary to establish an upper limit for the uncertainty and select data whose uncertainty is less than the limit. Selection of data that has one or more degrees of freedom is more complicated. It is necessary not only to consider the uncertainty limit but also on the way the data are distributed over state variable space. An effective algorithm has been developed by the Thermodynamics Research Center and used several data evaluations. It is described in several published references [1,2,4] and has also been used in an extensive review of virial coefficients now in press.

Briefly, the selection of each particular property value is based on a comparison of the uncertainty for the property with a weighed mean of uncertainties of neighboring properties. The weighting factor is an inverse exponential function of the difference between the state variables of the property in question and the state variables of each neighboring point. The selection level and the parameters in the weighting function depend on the size of the data set, the range of state variables it covers, the kind of property being selected and other considerations.

Data with one or more degrees of freedom returned from Level 4 is now suitable for fitting to smoothing functions or theoretical models by the least squares criteria. Level 5 returns parameters for the selected model as well as statistical measures of the goodness of fit. Reciprocals of the square of the uncertainties in properties serves as weighting factors for the fit. Combination of the parameters with Level 4 data permits comparison of the selected property values with those predicted by the model.

Level 5 fits functions to the group of properties defined in Level 2 and returned from Level 4. This automatically guarantees internal consistency among these properties. However these results will not necessarily be thermodynamically consistent with other kinds of properties. To achieve a global internal consistency it is necessary to fit all related properties to a general P-V-T equation of state. Level 6 carries out multi-property fits of this kind by combining several sets of data from Level 4 retrievals.

The extent of experimental data may not be sufficient to support satisfactory operation of Level 5 or Level 6 functions. It may be then possible to supplement Level 4 results with data calculated by theoretical or empirical correlation techniques. The combined data sets can then be passed to Levels 5 and 6 functions. These are considered as Level 7 results. An example would be the combination of ideal gas thermodynamic functions calculated by partition functions based on molecular energy states by spectroscopy [5] with experimentally derived ideal gas heat capacity and entropy from the Level 4 function. Calculated ideal gas functions could also be included in the Level 6 step. Group additivity correlations are available for ideal gas functions [6], critical constants [7], and enthalpies of formation [8], among others.

5 IMPLEMENTATION OF THE DATAFETCH LIBRARY

At the present time DataFetch functions have been written and tested for all Level 1 retrievals and several Level 2 retrievals. Work is in progress on examples of Level 3 and 4 functions. These are written in the C++ language. Initial creation and testing is carried out on UNIX operating system. However since they do not contain any direct

user input/output porting to any system that supports a standard C++ compiler is routine. Data from the functions are returned as vectors of structures using container classes from the C++ Standard Template Library (STL).

The current suite of functions access a local version of the SOURCE database created by the c-tree Plus® library distributed by the FairCom Corporation. This is well-tested library of database functions that operate in either single or multi-user modes as well as in a server over a local network. It runs on any computer and operating system now in use. Object versions of these functions are incorporated in the DataFetch library and can be distributed to users without additional costs.

Since access to a database is kept in clearly recognized modules in the DataFetch functions they can be modified to access any version of the SOURCE database. This includes direct or Internet access to the working ORACLE version maintained at the Thermodynamics Research Center.

6 APPLICATION PROGRAMS

Library functions do not interact directly with users. User input/output is supported by application programs. Application programs that display results from DataFetch functions are included with the library. The outputs may be re-directed to files. Source files are included and demonstrate the interaction with the functions. The library also includes various utility programs that allow compound name or formula search for obtain registry numbers, and those that extract and display literature references in several formats.

An application program is planned to call the Level 5 retrieval and generate a local

database of parameters in smoothing functions. It can be designed for individual user requirements and then be used by other application programs. It could also be used for the auxiliary data required in the Level 3 normalization step. The organization and formatting of reports, conversion of units, production of graphics, etc. belong to application programs.

Users may write application programs that generate special reports or that carry out many kinds of scientific or technical computations. These could incorporate elaborate calculations or graphical displays. Most user-generated application programs would call DataFetch functions at Level 4 or higher.

REFERENCES

1. R. C. Wilhoit and K. N. Marsh, *Int. J. Thermophys.* **10**:247 (1999)

2. *Documentation for the TRC Source Database* (Thermodynamics Research Center, College Station TX, 77843 Nov. 1999), 207 pp. An abridged version can be downloaded from the TRC Web site, trcweb.tamu.edu.

3. M. Frenkel, Q. Dong, R. C. Wilhoit, and K Hall *TRC SOURCE Database: A Unique Tool for Automatic Production of Data Compilations*, Fourteenth Symposium on Thermophysical Properties, Boulder, Colorado, June 25, 2000

4. R. C. Wilhoit, K. N. Marsh, X. Hong, N. Gadala and M. Frenkel, *Landolt-Börnstein, Group IV. Physical Chemistry, Vol. 8, Thermodynamic Properties of Organic*

Compounds and Their Mixtures, Subvolume B. Densities of Aliphatic Hydrocarbons. Alkanes (Springer-Verlag, Berlin, 1996). Also Subvolumes C-F.

5. M. Frenkel, G. J. Kabo, K. N. Marsh, G. N. Roganov and R. C. Wilhoit, *Thermodynamics of Organic Compounds in the Gas State, Vols. I and II*, TRC Data Series, (CRC Press, 1994)

6. S. W. Benson, F. R., Cruickshank, D. M. Golden, H. E. Haugen, H. E. O'Neal. A. S. Rodgers, R. Shaw, and R. Walsh, *Chem. Rev.* **69**:279 (1969) and S. W. Benson, *Thermochemical Kinetics*, 2nd edition (John Wiley: New York, 1976)

7. A. L. Lyderson, *Engineering Experiment Station Report No. 3*, Univ. of Wisconsin (1955); K. G. Joback and R. C. Reid, *Chem. Eng. Commun.* **57**, 233 (1987); G. R. Somayajulu *J. Chem. Eng. Data* **34**, 106 (1989)

8. J. B. Pedley, *Thermochemical Data and Structures of Organic Compounds*, Vol. 1, TRC Data Series (CRC Press, 1994)